Data Science in R (DSR) – Testat 3

Markus Gumbel

05. Dezember 2023

Inhaltsverzeichnis

Hinweise

Aktion		Datum	Ablauf
Ausgabe	Do.,	05.12.23	über Moodle
Abgabe	Fr.,	05.01.24, 14:00 Uhr	über Moodle
Testat	Mo.,	08.01.24	als Präsentation

- Die Abnahme der Testate gilt als Prüfungsleistung. Bei einer Verhinderung durch Krankheit ist eine ärztliche Bescheinigung der Arbeitsunfähigkeit vorzulegen.
- Geben Sie Ihre Lösungen fristgerecht via Moodle ab. Es werden dort folgende Abgaben erwartet (Ausarbeitungen in einem anderen Format werden nicht berücksichtigt):
  - Für jede Aufgabe ein R-Notebook (als *.Rmd-Datei) und etwaige benötigten R-Skripte (*.R).
  - Genau ein PDF mit allen Lösungen, das außerdem die Gruppennummer und alle Namen enthält.
- Nutzen Sie das R-Notebook, das es für jede Aufgabe gibt und **fügen Sie dort Ihre Lösungen ein**, d.h. Source-Code und normaler Text.

- **Kommentieren Sie Ihre Lösungen ausführlich** – entweder als normalen Text im Notebook oder als Kommentar in R-Anweisungen. Ohne Kommentare gibt es Punktabzug.
- Jedes abgegebene R-Notebook muss ohne Fehler auf einem anderen Rechner mit R (4.1 oder höher) ausgeführt werden können und ein HTML-Dokument erzeugen. Installationsanweisungen für die benutzen Packages sind nicht nötig – laden Sie aber die Packages im Notebook.
- Während der Abnahme sind die Ergebnisse am Rechner live zu demonstrieren.
- Bei der Abnahme der Übung ist der Studentenausweis vorzulegen.

1 Vorhersage des IQ mit linearer Regression

In dieser Aufgabe soll mittels linearer Regression der Intelligenz-Quotient (IQ) von Kindern vorhergesagt werden.

1.1 Daten importieren und anschauen

Laden Sie die Daten unter https://oc.informatik.hs-mannheim.de/s/K2nJQngd8N6o3M5/download in einen Data Frame. Die Daten sind im RDS-Format gespeichert. Der Datensatz stammt von Martin Brand (brandt@psychologie.uni-mannheim.de, https://www.sowi.uni-mannheim.de/kuhlmann/team/akademische-mitarbeiterinnen-und-mitarbeiter/brandt-martin/). Die Spalten bedeuten:

- `id`: Eindeutige Nummer
- `IQ.mother`: IQ der Mutter
- `IQ.father`: IQ des Vaters
- `n.siblings`: Anzahl der Geschwister
- `IQ.child`: IQ des Kindes

```
# Ihre Lösung:
```

Verschaffen Sie sich einen ersten Überblick über die Daten. Plotten Sie dazu ein Mehrfach-Diagramm, das Scatterplots aller vier Variablen (drei Merkmale und Response `IQ.child`) untereinander zeigt. D.h. jede Variable soll mit jeder anderen paarweise verglichen werden.

Tipp: Mit der `caret::featurePlot`-Funktion lässt sich das gut erreichen.

```
# Ihre Lösung:
```

1.2 Lineare Regression

Wenden Sie für diesen Datensatz eine lineare Regression an, um den IQ von Kindern in Abhängigkeit der Merkmale schätzen zu können. Die lineare Regression soll im Objekt `lm.iq` gespeichert werden.

```
# Ihre Lösung:
```

1.2.1 Merkmale

Welche Merkmale spielen eine Rolle? Interpretieren Sie das Ergebnis.

```
# Ihre Lösung:
```

1.2.2 RSE

Berechnen Sie den RSE. Was gibt dieser an?

```
# Ihre Lösung:
```

1.2.3 Anteil erklärter Varianz

Wie groß (in Prozent) ist der Anteil der erklärten Varianz? Ist das Ergebnis zufriedenstellend?

```
# Ihre Lösung:
```

1.3 Geschwister-Anzahl als Faktor

1.3.1 Anwendung

Wandeln Sie den Datensatz so um, dass die Anzahl der Geschwister (`n.siblings`) ein Faktor ist und wenden Sie darauf die lineare Regression an.

```
# Ihre Lösung:
```

1.3.2 R^2

R^2 hat sich verändert. Interpretieren bzw. begründen Sie das Ergebnis:

1.4 Schätzung für Testdaten

Abschließend soll mittels Cross-Validation überprüft werden, wie gut dieses Verfahren für unbekannte Testdaten funktioniert. Überlegen Sie sich ein passendes Verfahren und bestimmen Sie damit die R^2-Statistik.

```
# Ihre Lösung:
```

2 Vorhersage des IQ mit KNN-Regression

In dieser Aufgabe soll mittels KNN-Regression der Intelligenz-Quotient (IQ) von Kindern vorhergesagt werden (siehe auch andere Aufgabe).

2.1 KNN-Regression erläutern

k-nearest-neighbors (KNN) ist ein sehr einfaches Verfahren, mit dem Daten klassifiziert oder eine Regression berechnet werden kann. Im Fall der Regression werden für einen Datensatz die nächsten $k \in \mathbb{N}$ Nachbarn ermittelt und der Mittelwert dieser k Response-Variablen berechnet – das ist der geschätzte Wert. Üblicherweise wird die euklidische Distanz zum Ermitteln der Nachbarn genutzt.

Ein Einführungsvideo finden Sie unter https://youtu.be/sTJApaBjong. Weitere Infos auch unter https://bookdown.org/tpinto_home/Regression-and-Classification/k-nearest-neighbours-regression.html

Beschreiben Sie mit eigenen Worten (180 bis 220 Wörter), wie dieses Verfahren für die Regression und Klassifikation angewandt werden kann. Was sind die Vor- und Nachteile? Was sind die Besonderheiten im Vergleich zu anderen Verfahren?

(Ihre Lösung als Text hier)

2.2 Daten importieren

Laden Sie abermals den Datensatz unter https://oc.informatik.hs-mannheim.de/s/K2nJQngd8N6o3M5/download in einen Data Frame.

```
# Ihre Lösung:
```

2.3 KNN-Regression anwenden

2.3.1 Anwendung

Wenden Sie für diesen Datensatz eine KNN-Regression an, um den IQ von Kindern in Abhängigkeit der Merkmale schätzen zu können. Im Package `caret` gibt es die Lernmethode `knn`, die dafür genutzt werden kann.

```
# Ihre Lösung:
```

2.3.2 k?

Finden Sie heraus, welcher Wert für die Anzahl der nächsten Nachbarn am besten funktioniert. Beschreiben Sie, wie Sie das gemacht haben.

(Ihre Lösung als Text oder Chunk hier)

2.3.3 R^2

Lassen Sie die R^2-Statistik berechnen. Wie sieht diese im Vergleich zur linearen Regression aus? Welches Verfahren ist besser?

(Ihre Lösung als Text hier)

```
# Ihre Lösung:
```

3 Package-Erstellung

In dieser Aufgabe soll ein R-Package erstellt werden, das lokal in einer anderen R-Umgebung installiert und ausgeführt werden kann. Sie dürfen sich selbst überlegen, was Ihr Paket machen soll – gerne etwas Sinnvolles. Die Randbedingungen lauten:

- Mindestens zwei exportierte Funktionen.
- Mindestens eine exportierte Funktion liefert ein *echtes* R-Objekt, d.h. eine Instanz einer S3- oder R6-Klasse.
- Mindestens eine interne Funktion, die nicht nach dem Laden des Packages sichtbar ist.
- Es soll etwas mit Statistik oder Data Science zu tun haben und auf Data Frames (siehe auch unten) angewandt werden können.

Neben den o.g. Dingen gehen das erfolgreiche Installieren und die nachfolgenden Punkte in die Bewertung ein.

3.1 Dokumentation

Schreiben Sie eine ausführliche Dokumentation:

- Was macht das Package allgemein?
- Was machen die Funktionen?
- Geben Sie Beispiele zur Anwendung in den Hilfeseiten.

Der Text kann in Deutsch oder Englisch geschrieben werden. Die Hilfeseiten sollen nach der Installation des Packages in RStudio lesbar sein.

3.2 Tests

Schreiben Sie mindestens 10 Testfälle mit `Testthat`, die Ihre Funktionen testen. Diese müssen alle erfolgreich sein, wenn das Package veröffentlicht wird.

3.3 Beispieldatensatz

Fügen Sie Ihrem Package einen kleinen Beispieldatensatz bei. Klein meint: ab 50 bis maximal 500 Datensätze.

3.4 Abgabe

Geben den Source-Code Ihres Packages als Zip-Datei ab und laden Sie das Package selbst hoch.